DOI:
10.1214/16-AOS1533

Document Version
Publisher's PDF, also known as Version of record

The Annals of Statistics
2017, Vol. 45, No. 6, 1–26
https://doi.org/10.1214/16-AOS1533

ADAPTIVE BERNSTEIN–VON MISES THEOREMS IN GAUSSIAN WHITE NOISE[1]

BY KOLYAN RAY

Leiden University

We investigate Bernstein–von Mises theorems for adaptive nonparametric Bayesian procedures in the canonical Gaussian white noise model. We consider both a Hilbert space and multiscale setting with applications in L^2 and L^∞, respectively. This provides a theoretical justification for plug-in procedures, for example the use of certain credible sets for sufficiently smooth linear functionals. We use this general approach to construct optimal frequentist confidence sets based on the posterior distribution. We also provide simulations to numerically illustrate our approach and obtain a visual representation of the geometries involved.

1. Introduction. A key aspect of statistical inference is uncertainty quantification and the Bayesian approach to this problem is to use the posterior distribution to generate a *credible set*, that is, a region of prescribed posterior probability (often 95%). This can be considered an advantage of the Bayesian approach since Bayesian credible sets can be computed by simulation. In particular, the Bayesian generates a number of posterior draws and then keeps a prescribed fraction of the draws, discarding the remainder which are considered "extreme" in some sense. From a frequentist perspective, key questions are whether such a method has a theoretical justification and what is an effective rule for determining which draws to discard. A natural approach is to characterize such draws using a geometric notion, in particular by considering a minimal ball in some metric.

In finite dimensions, the Euclidean distance has a clear interpretation as the natural measure of size. However, in infinite dimensions such a notion is less clear-cut: the L^2 metric is the natural generalization of the Euclidean norm, but lacks a clear visual interpretation, while L^∞ can be easily visualized but is more difficult to treat mathematically. From the Bayesian perspective of simulating credible sets, the practitioner ultimately seeks a practical and effective rule for sorting through posterior draws and such geometric interpretations can be viewed as somewhat artificial impositions. The aim of this article is therefore to study possible geometric choices of credible sets that behave well from a frequentist asymptotic perspective.

Received July 2015; revised December 2016.

[1]Supported by UK Engineering and Physical Sciences Research Council (EPSRC) Grant EP/H023348/1 and the European Research Council under ERC Grant Agreement 320637.

MSC2010 subject classifications. Primary 62G20; secondary 62G15, 62G08.

Key words and phrases. Bayesian inference, posterior asymptotics, adaptation, credible set, confidence set.

Consider data $Y^{(n)}$ arising from some probability distribution $\mathbb{P}_f^{(n)}$, $f \in \mathcal{F}$. We place a prior distribution Π on $\mathcal{F}$ and study the behaviour of the posterior distribution $\Pi(\cdot \mid Y^{(n)})$ under the frequentist assumption $Y^{(n)} \sim \mathbb{P}_{f_0}^{(n)}$ for some nonrandom true $f_0 \in \mathcal{F}$ as the data size or quality $n \to \infty$. From such a viewpoint, the theoretical justification for posterior based inference using any (Borel) credible set in finite dimensions is provided by the Bernstein–von Mises (BvM) theorem (see [30, 45]). This deep result establishes mild conditions on the prior under which the posterior is approximately a normal distribution centered at an efficient estimator of the true parameter. It thus provides a powerful tool to study the asymptotic behaviour of Bayesian procedures and justifies the use of Bayesian simulations for uncertainty quantification.

A BvM in infinite-dimensions fails to hold in even very simple cases. Freedman [18] showed that in the basic conjugate ℓ_2 sequence space setting with both Gaussian priors and data, the BvM does not hold for ℓ_2-balls centered at the posterior mean; see also the related contributions [16, 25, 29]. The resulting message is that despite their intuitive interpretation, credible sets based on posterior draws using an ℓ_2-based selection procedure do not behave as in classical parametric models. Recently, Castillo and Nickl [11, 12] have established fully infinite-dimensional BvMs by considering weaker topologies than the classical L^p spaces. Their focus lies on considering spaces which admit $1/\sqrt{n}$-consistent estimators and where Gaussian limits are possible, unlike L^p-type loss. Credible regions selected using these different geometries are shown to behave well, generating asymptotically exact frequentist confidence sets. In this paper, we explore this approach in practice via both theoretical results for *adaptive* priors, as well as by numerical simulations. We consider an empirical Bayes, a hierarchical Bayes and a multiscale Bayes approach.

This approach is numerically illustrated in Section 6, where adaptive credible sets from various geometries are obtained by simulation. The main message of these numerical examples is that simulating credible sets from these slightly different geometries yields sets that do not look particularly strange in practice, and in fact often resemble more "classical" credible sets. Both approaches are methodologically similar; the only difference being the rule for discarding posterior draws. From a theoretical point of view, the difference between the two approaches is far more significant, with one yielding exact coverage statements at the expense of unbounded diameter. It is however possible to improve upon the naive implementation of such sets to also obtain the optimal diameter (see Proposition 1 of [12] and related results below). Modifying the geometry in such a way to obtain an exact coverage statement therefore comes at little additional cost from a practitioner's perspective.

Nonparametric priors typically contain tuning or hyper parameters, and it is a key challenge to study procedures that select these parameters automatically in a data-driven manner. This avoids the need to make unreasonably strong prior assumptions about the unknown parameter of interest, since incorrect calibration of

the prior can lead to suboptimal performance (see, e.g., [27]). It therefore makes sense to use an automatic procedure, unless a practitioner is particularly confident that their prior correctly captures the fine details of the unknown parameter, such as its level of smoothness or regularity. Adaptive procedures are widely used in practice, with hyper parameters commonly selected using a hyperprior or an empirical Bayes method. In the case of Gaussian white noise, a number of Bayesian procedures have been shown to be rate adaptive over common smoothness classes (e.g., [24, 26, 37]). Most such frequentist analyses restrict attention to obtaining contraction rates and do not study coverage properties of credible sets. The focus of this paper is therefore to investigate nonparametric BvMs for adaptive priors, with the goal of studying the coverage properties of credible sets.

In the case of Gaussian white noise, there has been recent work [27, 29] circumventing the need for a BvM by explicitly studying the coverage properties of certain specific credible sets. Of particular relevance is a nice recent paper by Szabó et al. [43], where the authors use an empirical Bayes approach combined with scaling up the radius of ℓ_2-balls to obtain adaptive confidence sets under a so-called *polished tail condition*. Their approach relies on explicit prior computations and provides an alternative to the more abstract point of view taken here. One of our principal goals is exact coverage statements and this seems more difficult to obtain using such an explicit approach. Since adaptive confidence sets do not exist in full generality, we also require self-similarity conditions on the true parameter to exclude certain "difficult" functions [6, 21, 23]. In particular, we shall consider the procedure of [43] in Section 3.1 and obtain exact coverage statements under the self-similarity condition introduced there.

We note other work dealing with BvM results in the nonparametric setting. Leahu [29] has studied the impact of prior smoothness on the existence of BvM theorems in the conjugate Gaussian sequence space model. Bickel and Kleijn [3], Castillo [8], Rivoirard and Rousseau [41] and Castillo and Rousseau [13] provide sufficient conditions for BvMs for semiparametric functionals. For the case of finite-dimensional posteriors with increasing dimension, see Ghosal [19] and Bontemps [4] for the case of regression or Boucheron and Gassiat [5] for discrete probability distributions.

Much of the approach taken here can equally be applied to other statistical settings such as sparsity and inverse problems [38], but we restrict to the nonparametric regime for ease of exposition. Since our focus lies on BvM results and coverage statements and this changes little conceptually, we omit such generalizations to maintain mathematical clarity.

2. Statistical setting.

2.1. *Function spaces and the white noise model.* We use the usual notation $L^p = L^p([0,1])$ for p-times Lebesgue integrable functions and denote by ℓ_p the usual sequence spaces. We consider the canonical white noise model, which is

equivalent to the fixed design Gaussian regression model with known variance. For $f \in L^2 = L^2([0,1])$, consider observing the trajectory:

$$dY_t^{(n)} = f(t)\,dt + \frac{1}{\sqrt{n}}\,dB_t, \qquad t \in [0,1], \tag{2.1}$$

where dB is a standard white noise. By considering the action of an orthonormal basis $\{e_\lambda\}_{\lambda\in\Lambda}$ on (2.1), it is statistically equivalent to consider the Gaussian sequence space model:

$$Y_\lambda^{(n)} \equiv Y_\lambda = f_\lambda + \frac{1}{\sqrt{n}} Z_\lambda, \qquad \lambda \in \Lambda, \tag{2.2}$$

where the $(Z_\lambda)_{\lambda\in\Lambda}$ are i.i.d. standard normal random variables and the unknown parameter of interest $f = (f_\lambda)_{\lambda\in\Lambda}$ is assumed to be in ℓ_2. We denote by $\mathbb{P}_{f_0}$ or $\mathbb{P}_0$ the law of Y arising from (2.2) under the true function f_0. In the following, Λ will represent either a Fourier-type basis or a wavelet basis. In the ℓ_2-setting, (2.2) can be interpreted purely in sequence form with $\Lambda = \mathbb{N}$ and we do not need to associate to it a time index $t \in [0,1]$.

In L^∞, we consider a multiscale approach so that $\Lambda = \{(j,k) : j \geq 0, k = 0, \ldots, 2^j - 1\}$. In particular, we consider an S-regular ($S \geq 0$) wavelet basis of $L^2([0,1])$, $\{\psi_{lk} : l \geq J_0 - 1, k = 0, \ldots, 2^l - 1\}$, with $J_0 \in \mathbb{N}$. For notational simplicity, denote the scaling function ϕ by the first wavelet $\psi_{(J_0-1)0}$. We consider either periodized wavelets or boundary corrected wavelets (see [33] for more details). Moreover, in certain applications we require in addition that the wavelets satisfy a localization property:

$$\begin{aligned} \sup_{x\in[0,1]} \sum_{k=0}^{2^{J_0}-1} \bigl|\phi_{J_0k}(x)\bigr| &\leq c(\phi) 2^{J_0/2} < \infty, \\ \sup_{x\in[0,1]} \sum_{k=0}^{2^{j}-1} \bigl|\psi_{jk}(x)\bigr| &\leq c(\psi) 2^{j/2} < \infty, \end{aligned} \tag{2.3}$$

$j \geq J_0$ (see Section 8.3 in the Supplement [39] for more discussion). The sequence model (2.2) corresponds to estimating the wavelet coefficients $f_{lk} = \langle f, \psi_{lk}\rangle$, for all $(l,k) \in \Lambda$, since any function $f \in L^2$ generates such a wavelet sequence. Conversely, any such sequence (f_{lk}) generates the wavelet series of a function (or distribution if the sequence is not in ℓ_2) $\sum_{(l,k)} f_{lk}\psi_{lk}$.

For $s, \delta \geq 0$, define the Sobolev spaces at the logarithmic level:

$$H^{s,\delta} \equiv H_2^{s,\delta} := \left\{ f \in \ell_2 : \|f\|_{s,2,\delta}^2 := \sum_{k=1}^{\infty} k^{2s} (\log k)^{-2\delta} |f_k|^2 < \infty \right\}.$$

From this, we recover the usual definition of the Sobolev spaces $H^s \equiv H_2^s = H_2^{s,0}$ and by duality we define for $s > 0$, $H_2^{-s} := (H_2^s)^*$. By standard Hilbert space

duality arguments, we can consider ℓ_2 as a subspace of H_2^{-s} and can similarly define the logarithmic spaces for $s < 0$ and $\delta \geq 0$ using the above series definition. In the ℓ_2-setting, we shall classify smoothness via the Sobolev *hyper rectangles* for $\beta \geq 0$:

$$\mathcal{Q}(\beta, R) = \Big\{ f \in \ell_2 : \sup_{k \geq 1} k^{2\beta+1} f_k^2 \leq R \Big\}.$$

In the $L^\infty([0,1])$-setting, we consider multiscale spaces: for a monotone increasing sequence $w = (w_l)_{l \geq 1}$ with $w_l \geq 1$, define

$$\mathcal{M} = \mathcal{M}(w) = \left\{ x = (x_{lk}) : \|x\|_{\mathcal{M}(w)} := \sup_{l \geq 0} \frac{1}{w_l} \max_k |x_{lk}| < \infty \right\}$$

(for further references to multiscale statistics see [12]). A separable closed subspace is obtained by considering the restriction:

$$\mathcal{M}_0 = \mathcal{M}_0(w) = \left\{ x \in \mathcal{M}(w) : \lim_{l \to \infty} \frac{1}{w_l} \max_k |x_{lk}| = 0 \right\},$$

that is those (weighted) sequences in $\mathcal{M}(w)$ that converge to 0. Note that $\mathcal{M}$ contains the space ℓ_2, since $\|x\|_{\mathcal{M}} \leq \|x\|_{\ell_2}$ as $w_l \geq 1$. In this setting, we consider norm-balls in the Besov spaces $B^\beta_{\infty\infty}([0,1])$:

$$\mathcal{H}(\beta, R) = \big\{ f = (f_{lk})_{(l,k) \in \Lambda} : |f_{lk}| \leq R 2^{-l(\beta+1/2)}, \forall (l,k) \in \Lambda \big\}.$$

We recall that $B^\beta_{\infty\infty}([0,1]) = C^\beta([0,1])$, the classical Hölder (–Zygmund in the case $\beta \in \mathbb{N}$) spaces. For more details on these embeddings and identifications, see [33].

Whether an ℓ_2-white noise defines a tight random element of $\mathcal{M}_0(w)$ depends on the weighting sequence (w_l). Recall that we call a sequence $(w_l)_{l \geq 1}$ *admissible* if $w_l/\sqrt{l} \nearrow \infty$ as $l \to \infty$ [12]. Let $Z = \{Z_\lambda = \langle Z, e_\lambda \rangle : \lambda \in \Lambda\}$, where $Z_\lambda \sim N(0,1)$ i.i.d., denote the Gaussian white noise in (2.2). We have from [11, 12] that for $\delta > 1/2$ and (w_l) an admissible sequence, Z defines a tight Gaussian Borel random variable on $H_2^{-1/2,\delta}$ and $\mathcal{M}_0(w)$, respectively, which we denote $\mathbb{Z}$. In view of this tightness, we can consider (2.1) as a Gaussian shift model:

$$\mathbb{Y}^{(n)} = f + \frac{1}{\sqrt{n}} \mathbb{Z},$$

where the above inequality is in the $H_2^{-1/2,\delta}$- or $\mathcal{M}_0(w)$-sense. Since $\sqrt{n}(\mathbb{Y}^{(n)} - f) = \mathbb{Z}$ in $H_2^{-1/2,\delta}$ or $\mathcal{M}_0(w)$, it immediately follows that $\mathbb{Y}^{(n)}$ is an efficient estimator of f in either norm.

Among the two classes $\{H_2^{s,\delta}\}_{s \in \mathbb{R}, \delta \geq 0}$ and $\{\mathcal{M}_0(w)\}_w$ of spaces considered, one can show that $s = -1/2$, $\delta > 1/2$ and admissibility of w determine the minimal spaces where the law of the ℓ_2-white noise Z is tight (see [11, 12] for further discussion). We therefore focus attention on these spaces since they provide the

threshold for which a weak convergence approach can work. For convenience, we denote $H \equiv H(\delta) \equiv H_2^{-1/2,\delta}$. We further denote the law of $\mathbb{Z}$ in H or $\mathcal{M}_0(w)$ by $\mathcal{N}$ as appropriate.

2.2. *Weak Bernstein–von Mises phenomena.* Due to the continuous embeddings $\ell_2 \subset H$ and $\ell_2 \subset \mathcal{M}_0(w)$, any Borel probability measure on ℓ_2 yields a tight Borel probability measure on H and $\mathcal{M}_0(w)$. Consider a prior Π on ℓ_2 and let $\Pi_n = \Pi(\cdot \mid Y^{(n)})$ denote the posterior distribution based on data (2.2). For S a vector space and $z \in S$, consider the map $\tau_z : S \to S$ given by

$$\tau_z : f \mapsto \sqrt{n}(f - z).$$

Let $\Pi_n \circ \tau_{\mathbb{Y}^{(n)}}^{-1}$ denote the image measure of the posterior distribution [considered as a measure on H or $\mathcal{M}_0(w)$] under the map $\tau_{\mathbb{Y}^{(n)}}$. Thus for any Borel set B arising from these topologies,

$$\Pi_n \circ \tau_{\mathbb{Y}^{(n)}}^{-1}(B) = \Pi\big(\sqrt{n}\big(f - \mathbb{Y}^{(n)}\big) \in B \mid Y^{(n)}\big),$$

so that we can more intuitively write $\Pi_n \circ \tau_{\mathbb{Y}^{(n)}}^{-1} = \mathcal{L}(\sqrt{n}(f - \mathbb{Y}^{(n)}) \mid Y^{(n)})$, where $\mathcal{L}(f \mid Y^{(n)})$ denotes the law of f under the posterior. For convenience, we metrize the weak convergence of probability measures via the bounded Lipschitz metric (defined in Section 8.4 in the Supplement [39]). Recalling that we denote by $\mathcal{N}$ the law of the white noise Z in (2.2) as an element of S, we define the notion of nonparametric BvM.

DEFINITION 1. Consider data generated from (2.2) under a fixed function f_0 and denote by $\mathbb{P}_0$ the distribution of $Y^{(n)}$. Let β_S be the bounded Lipschitz metric for weak convergence of probability measures on S. We say that a prior Π satisfies a weak Bernstein–von Mises phenomenon in S if, as $n \to \infty$,

$$\mathbb{E}_0 \beta_S\big(\Pi_n \circ \tau_{\mathbb{Y}^{(n)}}^{-1}, \mathcal{N}\big) = \mathbb{E}_0 \beta_S\big(\mathcal{L}\big(\sqrt{n}\big(f - \mathbb{Y}^{(n)}\big) \mid Y^{(n)}\big), \mathcal{N}\big) \to 0.$$

Here, S is taken to be one of $H(\delta)$ for $\delta > 1/2$, H^{-s} for $s > 1/2$ or $\mathcal{M}_0(w)$ for $(w_l)_{l \geq 1}$ an admissible sequence.

The weak BvM says that the (scaled and centered) posterior distribution asymptotically looks like an infinite-dimensional Gaussian distribution in some 'weak' sense, quantified via the bounded Lipschitz metric. Weak convergence in S implies that these two probability measures are approximately equal on certain classes of sets, whose boundaries behave smoothly with respect to the measure $\mathcal{N}$ (see Sections 1.1 and 4.1 of [11]).

2.3. *Self-similarity.* The study of adaptive BvM results naturally leads to the topic of adaptive frequentist confidence sets. It is known that confidence sets with

radius of optimal order over a class of submodels nested by regularity that also possess honest coverage do not exist in full generality (see [23, 35] for recent references). We therefore require additional assumptions on the parameters to be estimated and so consider self-similar functions, whose regularity is similar at both small and large scales. Such conditions have been considered in Giné and Nickl [21], Hoffmann and Nickl [23] and Bull [6] and ensure that we remove those functions whose norms (measuring smoothness) are difficult to estimate and which statistically look smoother than they actually are. We firstly consider the ℓ_2-type self-similarity assumption found in Szabó et al. [43].

DEFINITION 2. Fix an integer $N_0 \geq 2$ and parameters $\rho > 1$, $\varepsilon \in (0, 1)$. We say that a function $f \in \mathcal{Q}(\beta, R)$ is *self-similar* if

$$\sum_{k=N}^{\lceil \rho N \rceil} f_k^2 \geq \varepsilon R N^{-2\beta} \qquad \text{for all } N \geq N_0.$$

We denote the class of self-similar elements of $\mathcal{Q}(\beta, R)$ by $\mathcal{Q}_{SS}(\beta, R, \varepsilon)$.

This condition says that each block $(f_N, \ldots, f_{\lceil \rho N \rceil})$ of consecutive components contains at least a fixed fraction (in the ℓ_2-sense) of the size of a "typical" element of $\mathcal{Q}(\beta, R)$, so that the signal looks similar at all frequency levels (see [34, 35, 43] for further discussion). The parameters N_0 and ρ affect the results of this article through the sample size at which the asymptotic results take effect, that is, $n \to \infty$ implicitly implies statements of the form "for $n \geq n_0$ large enough", where n_0 depends on N_0 and ρ. For this reason, the impact of N_0 and ρ is not explicitly mentioned below and one may simply treat these constants as fixed (e.g., $N_0 = 2$ and $\rho = 2$). The lower bound in Definition 2 can be slightly weakened to permit, for example, logarithmic deviations from $N^{-2\beta}$. However, since this results in additional technicality whilst adding little extra insight, we do not pursue such a generalization here. It is possible to consider a weaker self-similarity condition using a strictly frequentist approach [35], though this has not been explored in the Bayesian setting and it is unclear whether our approach extends in such a way. Let $K_j(f) = \sum_k \langle f, \phi_{jk} \rangle \phi_{jk}$ denote the wavelet projection at resolution level j. In L^∞, we consider Condition 3 of Giné and Nickl [21], which can only be slightly relaxed [6].

DEFINITION 3. Fix a positive integer j_0. We say that a function $f \in \mathcal{H}(\beta, R)$ is *self-similar* if there exists a constant $\varepsilon > 0$ such that

$$\|K_j(f) - f\|_\infty \geq \varepsilon 2^{-j\beta} \qquad \text{for all } j \geq j_0.$$

We denote the class of self-similar elements of $\mathcal{H}(\beta, R)$ by $\mathcal{H}_{SS}(\beta, R, \varepsilon)$.

In particular, since $f \in \mathcal{H}(\beta, R)$, we have that $\|K_j(f) - f\|_\infty \asymp 2^{-j\beta}$ for all $j \geq j_0$. What we really require is that there is at least one significant coefficient at

the level $\log_2((n/\log n)^{1/(2\beta+1)})$ that the posterior distribution can detect. However, this level depends also on unknown constants in practice (see proof of Proposition 4.5) and so we require a statement for all (sufficiently large) resolution levels as in Definition 3. See Giné and Nickl [21] and also Bull [6] for further discussion about this condition.

3. Bernstein–von Mises results.

3.1. *Empirical and hierarchical Bayes in ℓ_2.* We continue the frequentist analysis of the adaptive priors studied in [26, 42, 43] in ℓ_2. For $\alpha > 0$, define the product prior on the ℓ_2-coordinates by the product measure

$$\Pi_\alpha = \bigotimes_{k=1}^{\infty} N\big(0, k^{-2\alpha-1}\big),$$

so that the coordinates are independent. A draw from this distribution will be Π_α-almost surely in all Sobolev spaces $H_2^{\alpha'}$ for $\alpha' < \alpha$. The posterior distribution corresponding to Π_α is given by

$$\Pi_\alpha(\cdot \mid Y) = \bigotimes_{k=1}^{\infty} N\left(\frac{n}{k^{2\alpha+1}+n} Y_k, \frac{1}{k^{2\alpha+1}+n}\right). \tag{3.1}$$

If $f_0 \in H^\beta$ and $\alpha = \beta$, it has been shown [2, 7, 27] that the posterior contracts at the minimax rate of convergence, while if $\alpha \neq \beta$, then strictly suboptimal rates are achieved. Since the true smoothness β is generally unknown, two data-driven procedures have been considered in [26]. The empirical Bayes procedure consists of selecting the smoothness parameter by using a likelihood-based approach. Namely, we consider the estimate

$$\hat{\alpha}_n = \operatorname*{argmax}_{\alpha \in [0, a_n]} \ell_n(\alpha), \tag{3.2}$$

where $a_n \to \infty$ is any sequence such that $a_n = o(\log n)$ as $n \to \infty$ and

$$\ell_n(\alpha) = -\frac{1}{2} \sum_{k=1}^{\infty} \left(\log\left(1 + \frac{n}{k^{2\alpha+1}}\right) - \frac{n^2}{k^{2\alpha+1}+n} Y_k^2 \right)$$

is the marginal log-likelihood for α in the joint model (f, Y) in the Bayesian setting [relative to the infinite product measure $\bigotimes_{k=1}^{\infty} N(0,1)$]. The quantity a_n is needed to uniformly control the finite dimensional projections of the empirical Bayes procedure to establish a parametric BvM (Theorem 7.2). The posterior distribution is defined via the plug-in procedure:

$$\Pi_{\hat{\alpha}_n}(\cdot \mid Y) = \Pi_\alpha(\cdot \mid Y)|_{\alpha = \hat{\alpha}_n}.$$

If there exist multiple maxima to (3.2), then any of them can be selected.

A fully Bayesian approach is to put a hyperprior on the parameter α. This yields the hierarchical prior distribution:

$$\Pi^H = \int_0^\infty \lambda(\alpha)\Pi_\alpha \, d\alpha,$$

where λ is a positive Lebesgue density on $(0, \infty)$ satisfying the following assumption (Assumption 1 of [26]).

CONDITION 1. *Assume that for every $c_1 > 0$, there exists $c_2 \geq 0$, $c_3 \in \mathbb{R}$, with $c_3 > 1$ if $c_2 = 0$ and $c_4 > 0$ such that for $\alpha \geq c_1$,*

$$c_4^{-1}\alpha^{-c_3}\exp(-c_2\alpha) \leq \lambda(\alpha) \leq c_4\alpha^{-c_3}\exp(-c_2\alpha).$$

The exponential, gamma and inverse gamma distributions satisfy Condition 1 for example. Knapik et al. [26] showed that both these procedures contract to the true parameter adaptively at the (almost) minimax rate, uniformly over Sobolev balls, and a similar result holds for Sobolev hyper rectangles. Both procedures satisfy weak BvMs in the sense of Definition 1.

THEOREM 3.1. *Consider the empirical Bayes procedure described above. For every $\beta, R > 0$ and $s > 1/2$, we have*

$$\sup_{f_0 \in \mathcal{Q}(\beta,R)} \mathbb{E}_0 \beta_{H^{-s}}\big(\Pi_{\hat{\alpha}_n} \circ \tau_{\mathbb{Y}}^{-1}, \mathcal{N}\big) \to 0$$

as $n \to \infty$. Moreover, for $\delta > 2$ we have the (slightly) stronger convergence:

$$\sup_{f_0 \in \mathcal{Q}_{SS}(\beta,R,\varepsilon)} \mathbb{E}_0 \beta_{H(\delta)}\big(\Pi_{\hat{\alpha}_n} \circ \tau_{\mathbb{Y}}^{-1}, \mathcal{N}\big) \to 0$$

as $n \to \infty$.

THEOREM 3.2. *Consider the hierarchical Bayes procedure described above, where the prior density λ satisfies Condition 1. For every $\beta, R > 0$ and $s > 1/2$, we have*

$$\sup_{f_0 \in \mathcal{Q}(\beta,R)} \mathbb{E}_0 \beta_{H^{-s}}\big(\Pi_n^H \circ \tau_{\mathbb{Y}}^{-1}, \mathcal{N}\big) \to 0$$

as $n \to \infty$. Moreover, for $\delta > 2$ we have the (slightly) stronger convergence:

$$\sup_{f_0 \in \mathcal{Q}_{SS}(\beta,R,\varepsilon)} \mathbb{E}_0 \beta_{H(\delta)}\big(\Pi_n^H \circ \tau_{\mathbb{Y}}^{-1}, \mathcal{N}\big) \to 0$$

as $n \to \infty$.

The requirement of self-similarity for a weak BvM in $H(\delta)$ could conceivably be relaxed, but such an assumption is natural since it is anyway needed for the construction of adaptive confidence sets in Section 4.1. It is not clear whether this

is a fundamental limit or a technical artefact of the proof. The condition $\delta > 2$ is also required for technical reasons.

Whilst minimax optimality is clearly desirable from a theoretical frequentist perspective, it may be too stringent a goal in our context. Using a purely Bayesian point of view, we derive an analogous result to Doob's almost sure consistency result. Specifically, a weak BvM holds in $H(\delta)$ for prior draws, almost surely under both the empirical Bayes and hierarchical priors. For this, it is sufficient to show that prior draws are self-similar almost surely.

PROPOSITION 3.3. *The parameter f_0 is self-similar in the sense of Definition 2, Π_α-almost-surely for any $\alpha > 0$. Consequently, $\Pi_{\hat{\alpha}_n}$ and Π^H satisfy a weak BvM in $H(\delta)$ for $\delta > 2$, Π_α-a.s., $\alpha > 0$, and Π^H-a.s., respectively.*

In particular, f satisfies Definition 2 with smoothness α and parameters $\rho > 1$ and $\varepsilon = \varepsilon(\alpha, \rho, R) > 0$ sufficiently small and random N_0 sufficiently large, Π_α-almost surely. As a simple corollary to Theorems 3.1 and 3.2, we have that the rescaled posteriors merge weakly [with respect to weak convergence on $H(\delta)$] in the sense of Diaconis and Freedman [17]. By Proposition 2.1 of [36], we immediately have that the unscaled posteriors merge weakly with respect to the ℓ_2-topology since they are both consistent [26]. However, in the case of bounded Lipschitz functions (rather than the full case of continuous and bounded functions), we can improve this result to obtain a rate of convergence.

COROLLARY 3.4. *For every $\beta, R > 0$, $s > 1/2$ and $\delta > 2$, we have*

$$\sup_{f_0 \in \mathcal{Q}(\beta,R)} \mathbb{E}_0 \beta_{H^{-s}}\big(\Pi_n^H \circ \tau_{\mathbb{Y}}^{-1}, \Pi_{\hat{\alpha}_n} \circ \tau_{\mathbb{Y}}^{-1}\big) \to 0,$$

$$\sup_{f_0 \in \mathcal{Q}_{SS}(\beta,R,\varepsilon)} \mathbb{E}_0 \beta_{H(\delta)}\big(\Pi_n^H \circ \tau_{\mathbb{Y}}^{-1}, \Pi_{\hat{\alpha}_n} \circ \tau_{\mathbb{Y}}^{-1}\big) \to 0$$

as $n \to \infty$. In particular, for $S = H^{-s}$ or $H(\delta)$ as above,

$$\sup_{u:\|u\|_{BL} \le L} \left| \int_S u \, d\big(\Pi_n^H - \Pi_{\hat{\alpha}_n}\big) \right| = o_{\mathbb{P}_0}\left(\frac{L}{\sqrt{n}}\right).$$

3.2. *Slab and spike prior in L^∞.* Consider the slab and spike prior, whose frequentist contraction rate has been analyzed in Castillo and van der Vaart [15], Hoffmann et al. [24] and Castillo et al. [14]. The assumptions in [24] ensure that prior draws are very sparse and only very few coefficients are fitted. We therefore modify the prior slightly so that the prior automatically fits the first few coefficients of the signal without any thresholding. This ensures that the posterior will have a rough approximation of the signal before fitting wavelet coefficients more sparsely at higher resolution levels. This makes sense from a practical point of view by preventing overly sparse models and is in fact necessary from a theoretical perspective (see Proposition 3.7).

Let $J_n = \lfloor \log n / \log 2 \rfloor$ be such that $n/2 < 2^{J_n} \leq n$ and define some strictly increasing sequence $j_0 = j_0(n) \to \infty$ such that $j_0(n) < J_n$. For the low resolutions $j \leq j_0(n)$, we fit a simple product prior where we draw the f_{jk}'s independent from a bounded density g that is strictly positive on $\mathbb{R}$. For the middle resolution levels $j_0(n) < j \leq J_n$, the f_{jk}'s are drawn independently from the mixture

$$\Pi_j(dx) = (1 - w_{j,n})\delta_0(dx) + w_{j,n} g(x)\, dx, \qquad n^{-K} \leq w_{j,n} \leq 2^{-j(1+\theta)},$$

for some $K > 0$ and $\theta > 1/2$. All coefficients at levels $j > J_n$ are set to 0. Since this is a product prior, one can sample from the posterior by sampling from each component separately (using either an MCMC scheme or explicit expressions depending on the choice of density g). We have a weak BvM in the multiscale space $\mathcal{M}_0(w)$, where the rate at which the admissible sequence (w_l) diverges depends on the how many coefficients we automatically fit in the prior via the sequence $j_0(n)$. Recall that a sequence $(w_l)_{l \geq 1}$ is admissible if $w_l / \sqrt{l} \nearrow \infty$.

THEOREM 3.5. *Consider the slab and spike prior defined above with lower threshold given by the strictly increasing sequence $j_0(n) \to \infty$. The posterior distribution satisfies a weak BvM in $\mathcal{M}_0(w)$ in the sense of Definition 1, that is, for every $\beta, R > 0$,*

$$\sup_{f_0 \in \mathcal{H}(\beta, R)} \mathbb{E}_0 \beta_{\mathcal{M}_0(w)}\big(\Pi_n \circ \tau_{\mathbb{Y}}^{-1}, \mathcal{N}\big) \to 0$$

as $n \to \infty$, for any admissible sequence (w_l) satisfying $w_{j_0(n)} / \sqrt{\log n} \nearrow \infty$.

Note that in the limiting case $w_l = \sqrt{l}$, we recover $j_0(n) \simeq \log n$, so that the prior automatically fits the same fixed fraction of the full $2^{J_n} \simeq n$ coefficients. Since we consider only admissible sequences, the fraction of coefficients that the prior fits automatically is asymptotically vanishing. An alternative way to consider this result is in reverse: based on a desired rate in practice, we prescribe an admissible sequence $w_l = \sqrt{l} u_l$, where u_l is some divergent sequence, and then pick $j_0(n)$ appropriately. Taking $j_0(n)$ to grow more slowly than any power of $\log n$ means (w_l) must grow faster than any power of l, resulting in a greater than logarithmic down-weighting of the wavelet coefficients in $\mathcal{M}(w)$. It may therefore be more appropriate to take $j_0(n)$ a power of $\log n$, which yields the following specific case.

COROLLARY 3.6. *Consider the slab and spike prior defined above with lower threshold $j_0(n) \simeq (\log n)^{\frac{1}{2\epsilon+1}}$ for some $\epsilon > 0$. Then it satisfies a weak BvM in $\mathcal{M}_0(w)$ in the sense of Definition 1, that is, for every $\beta, R > 0$,*

$$\sup_{f_0 \in \mathcal{H}(\beta, R)} \mathbb{E}_0 \beta_{\mathcal{M}_0(w)}\big(\Pi_n \circ \tau_{\mathbb{Y}}^{-1}, \mathcal{N}\big) \to 0$$

as $n \to \infty$ for the admissible sequence $w_l = l^{1/2+\epsilon} u_l$, where u_l is any (arbitrarily slowly) diverging sequence.

While the requirement to fit the first few coefficients of the prior is mild and of practical use in nonparametrics, it is naturally of interest to study the behaviour of the posterior distribution with full thresholding, that is, when $j_0(n) \equiv 0$, which we denote by Π'. In general, however, the full posterior contracts to the truth at a rate strictly slower than $1/\sqrt{n}$ in $\mathcal{M}(w)$, so that a $\sqrt{n}$-rescaling of the posterior cannot converge weakly to a limit. This holds even for self-similar functions.

PROPOSITION 3.7. *Let (w_l) be any admissible sequence. Then for any $\beta, R > 0$, there exists $\varepsilon = \varepsilon(\beta, R, \psi) > 0$ and $f_0 \in \mathcal{H}_{SS}(\beta, R, \varepsilon)$ such that along some subsequence (n_m),*

$$\mathbb{E}_0 \Pi'\big(\|f - \mathbb{Y}\|_{\mathcal{M}(w)} \geq M_{n_m} n_m^{-1/2} \mid Y^{(n_m)}\big) \to 1$$

for all $M_n \to \infty$ sufficiently slowly. Consequently, for such an f_0, a weak BvM in $\mathcal{M}_0(w)$ in the sense of Definition 1 cannot hold.

It is particularly relevant that Proposition 3.7 applies to self-similar parameters since a major application of the weak BvM is the construction of adaptive credible regions with good frequentist properties under self-similarity (see Proposition 4.5). On the level of a $\sqrt{n}$-rescaling as in Definition 1, the rescaled posterior distribution asymptotically puts vanishingly small probability mass on any given $\mathcal{M}(w)$-ball infinitely often. This occurs because the posterior selects nonzero coordinates by thresholding at the level $\sqrt{\log n/n}$ rather than the required $1/\sqrt{n}$ (Lemma 1 of [24]). The weighting sequence (w_l) regularizes the extra $\sqrt{\log n}$ factor at high frequencies, but does not do so at low frequencies. This is the reason that the weighting sequence (w_l) depends explicitly on the thresholding factor $\sqrt{\log n}$ in Theorem 3.5.

It seems that using such an adaptive scheme on low frequencies of the signal causes the weak BvM to fail. This prior closely resembles the frequentist practice of wavelet thresholding, where such a phenomenon has also been observed. For example, Giné and Nickl [20] require similar (though stronger) assumptions on the number of coefficients that need to be fitted automatically to obtain a central limit theorem for the distribution function of the hard thresholding wavelet estimator in density estimation (Theorem 8 of [20]).

4. Applications.

4.1. *Adaptive credible sets in ℓ_2.* We propose credible sets from the hierarchical or empirical Bayes procedures, which we show are adaptive frequentist confidence sets for self-similar parameters. We consider the natural Bayesian approach of using the quantiles of the posterior distribution to obtain a credible set of prescribed posterior probability. By considering sets whose geometry is amenable to the space $H(\delta)$, the weak BvM implies that such credible sets are asymptotically confidence sets.

Recall that $\|f\|_{H(\delta)}^2 = \sum_{k=1}^{\infty} k^{-1}(\log k)^{-2\delta} f_k^2$. For a given significance level $0 < \gamma < 1$, consider the credible set:

$$C_n = \{f : \|f - \mathbb{Y}\|_{H(\delta)} \le R_n/\sqrt{n}\}, \tag{4.1}$$

where $R_n = R_n(Y, \gamma)$ is chosen such that $\Pi_{\hat{\alpha}_n}(C_n|Y) = 1 - \gamma$ or $\Pi^H(C_n|Y) = 1 - \gamma$. Since the empirical and hierarchical Bayes procedures both satisfy a weak BvM in $H(\delta)$, we have from Theorem 1 of [11] that in both cases

$$\mathbb{P}_{f_0}(f_0 \in C_n) \to 1 - \gamma \quad \text{and} \quad R_n = O_{\mathbb{P}_0}(1)$$

as $n \to \infty$, so that C_n is asymptotically an exact frequentist confidence set (of unbounded ℓ_2-diameter). We control the diameter of the set using either the estimator $\hat{\alpha}_n$ or the posterior median as a smoothness estimate, and then use the standard frequentist approach of undersmoothing. In the first case, consider

$$\tilde{C}_n = \{f : \|f - \mathbb{Y}\|_{H(\delta)} \le R_n/\sqrt{n}, \|f - \hat{f}_n\|_{H^{\hat{\alpha}_n - \epsilon_n}} \le C\sqrt{\log n}\}, \tag{4.2}$$

where $\hat{f}_n$ is the posterior mean, R_n is chosen as in C_n, ϵ_n (chosen possibly data dependently) satisfies $r_1/(\log n) \le \epsilon_n \le (r_2/\log n) \wedge (\hat{\alpha}_n/2)$ for some $0 < r_1 \le r_2 \le \infty$ and $C > 1/r_1$. The undersmoothing by ϵ_n is necessary since the posterior assigns probability one to $H^{\alpha'}$ for $\alpha' < \hat{\alpha}_n$, while probability zero to $H^{\hat{\alpha}_n}$ itself. Geometrically, $\tilde{C}_n$ is the intersection of two ℓ_2-ellipsoids, C_n and an $H^{\hat{\alpha}_n - \epsilon_n}$-norm ball. For a typical element f in $\tilde{C}_n$, the size of the low frequency coordinates of f are determined by C_n, while the smoothness condition in $\tilde{C}_n$ acts to regularize the elements of C_n (which are typically not in ℓ_2) by shrinking the higher frequencies.

PROPOSITION 4.1. *Let $0 < \beta_1 \le \beta_2 < \infty$, $R \ge 1$ and $\varepsilon > 0$. Then the confidence set $\tilde{C}_n$ given in (4.2) satisfies*

$$\sup_{\substack{f_0 \in \mathcal{Q}_{SS}(\beta, R, \varepsilon) \\ \beta \in [\beta_1, \beta_2]}} \left|\mathbb{P}_{f_0}(f_0 \in \tilde{C}_n) - (1 - \gamma)\right| \to 0$$

as $n \to \infty$. For every $\beta \in [\beta_1, \beta_2]$, uniformly over $f_0 \in \mathcal{Q}_{SS}(\beta, R, \varepsilon)$,

$$\Pi_{\hat{\alpha}_n}(\tilde{C}_n \mid Y) = 1 - \gamma + O_{\mathbb{P}_0}\big(n^{-C' n^{1/(4\beta+2)}}\big)$$

for some $C' > 0$ independent of $\beta, R, \varepsilon, N_0, \rho$, while the ℓ_2-diameter satisfies for $\delta > 2$,

$$|\tilde{C}_n|_2 = O_{\mathbb{P}_0}\big(n^{-\beta/(2\beta+1)}(\log n)^{(2\delta\beta+1/2)/(2\beta+1)}\big).$$

The logarithmic correction in the definition of $H(\delta)$ that is required for a weak BvM causes the $(\log n)^{2\delta\beta/(2\beta+1)}$ penalty [which is $O((\log n)^{2\delta})$ uniformly over $\beta \ge 0$]; this is the price required for using a plug-in approach in $H(\delta)$. The remaining $(\log n)^{1/(4\beta+2)}$ factor arises due to the second constraint in $\tilde{C}_n$, where the

$H^{\hat{\alpha}_n}$-radius must be taken sufficiently large to ensure $\tilde{C}_n$ has sufficient posterior probability.

While the second constraint in (4.2) reduces the credibility below $1-\gamma$, Proposition 4.1 shows that this credibility loss is very small. The Bayesian approach takes care of this automatically since the posterior concentrates on a much more regular set than ℓ_2. This is corroborated empirically by numerical evidence (see Table 1), which shows that the credibility of the set $\tilde{C}_n$ rapidly approaches $1-\gamma$ as n increases.

REMARK 4.2. A naive interpretation of C_n yields a credible set that is far too large, having unbounded ℓ_2-diameter, with the additional constraint in $\tilde{C}_n$ needed to regularize the set. In actual fact, the posterior does this regularization automatically with C_n being "almost optimal". Proposition 4.1 could be rewritten for C_n with exact credibility $\Pi_{\hat{\alpha}_n}(C_n \mid Y) = 1-\gamma$ and ℓ_2-diameter satisfying

$$\begin{aligned}&\Pi_{\hat{\alpha}_n}\big(f \in C_n : \|f - \hat{f}_n\|_2 \le Cn^{-\frac{\beta}{2\beta+1}}(\log n)^{\frac{2\delta\beta+1/2}{2\beta+1}} \mid Y\big)\\ &\qquad = 1-\gamma + O_{\mathbb{P}_0}\big(n^{-C'n^{1/(4\beta+2)}}\big),\end{aligned}$$

for some $C, C' > 0$. In view of this, the sets C_n and $\tilde{C}_n$ are essentially the same from the point of view of the posterior, with C_n having exact credibility for finite n and correct ℓ_2-diameter asymptotically and $\tilde{C}_n$ having the reverse. In particular, the finite time credibility "gap" for either having too large radius in C_n or smaller than $1-\gamma$ credibility for $\tilde{C}_n$ is of the same size. Moreover, the above statement holds without the need for a self-similarity assumption, which is possible since the confidence set does not strictly have optimal diameter. The same notion also holds for C_n arising from the hierarchical Bayes procedure.

REMARK 4.3. By Lemma 8.2 in the Supplement [39], the empirical Bayes posterior mean $\hat{f}_n$ satisfies $\|\hat{f}_n - \mathbb{Y}\|_{H(\delta)} = o_{\mathbb{P}_0}(1/\sqrt{n})$ and so is an efficient estimator of f_0 in $H(\delta)$. Consequently, one can substitute $\mathbb{Y}$ with $\hat{f}_n$ in the definitions of C_n and $\tilde{C}_n$.

Replacing the estimate $\hat{\alpha}_n$ with the median α_n^M of the marginal posterior distribution $\lambda_n(\cdot|Y)$ yields a fully Bayesian analogue. To obtain the necessary undersmoothing over a target range $[\beta_1, \beta_2]$, we consider the shifted estimator $\hat{\beta}_n = \alpha_n^M - (C+1)/\log n$, where $C(R, \beta_2, \varepsilon, \rho) = \max_{\beta_1 \le \beta \le \beta_2} C(R, \beta, \varepsilon, \rho)$ is the constant appearing in Lemma 8.7 in the Supplement [39] (which can be explicitly computed). Consider

$$\tilde{C}_n' = \big\{f : \|f - \mathbb{Y}\|_H \le R_n/\sqrt{n}, \|f - \hat{f}_n\|_{H^{\hat{\beta}_n}} \le M_n\sqrt{\log n}\big\}, \tag{4.3}$$

where $\hat{f}_n$ is the posterior mean, $M_n \to \infty$ grows more slowly than any polynomial and R_n is chosen as in C_n. Taking C_n arising from the hierarchical Bayesian procedure Π^H, $\tilde{C}_n'$ is a "fully Bayesian" object. We have an analogue of Proposition 4.1.

PROPOSITION 4.4. *Let $0 < \beta_1 \le \beta_2 < \infty$, $R \ge 1$ and $\varepsilon > 0$. Then the confidence set $\tilde{C}_n'$ given in (4.3) satisfies*

$$\sup_{\substack{f_0 \in \mathcal{Q}_{SS}(\beta,R,\varepsilon) \\ \beta \in [\beta_1,\beta_2]}} \left| \mathbb{P}_{f_0}\left(f_0 \in \tilde{C}_n'\right) - (1-\gamma) \right| \to 0$$

as $n \to \infty$. For every $\beta \in [\beta_1, \beta_2]$, uniformly over $f_0 \in \mathcal{Q}_{SS}(\beta, R, \varepsilon)$,

$$\Pi^H\left(\tilde{C}_n' \mid Y\right) = 1 - \gamma + o_{\mathbb{P}_0}(1),$$

while the ℓ_2-diameter satisfies for $\delta > 2$,

$$\left|\tilde{C}_n'\right|_2 = O_{\mathbb{P}_0}\left(n^{-\beta/(2\beta+1)} (\log n)^{(2\delta\beta+1/2)/(2\beta+1)}\right).$$

4.2. *Adaptive credible bands in L^∞.* We provide a fully Bayesian construction of adaptive credible bands using the slab and spike prior. The posterior median $\tilde{f} = (\tilde{f}_{n,lk})_{(l,k)\in\Lambda}$ (defined coordinate-wise) takes the form of a thresholding estimator (cf. [1]), which we use to identify significant coefficients. This has the advantage of both simplicity and interpretability and also provides a natural Bayesian approach for this coefficient selection. Such an approach was used by Kueh [28] to construct an asymptotically honest (i.e., uniform in the parameter space) adaptive frequentist confidence set on the sphere using needlets. In that article, the coefficients are selected based on the empirical wavelet coefficients with the thresholds selected conservatively using Bernstein's inequality. In contrast, we use a Bayesian approach to automatically select the thresholding quantile constants that then yields exact coverage statements.

Let

$$D_n = \left\{ f : \|f - \mathbb{Y}\|_{\mathcal{M}(w)} \le R_n/\sqrt{n} \right\}, \tag{4.4}$$

where $R_n = R_n(Y, \gamma)$ is chosen such that $\Pi(D_n \mid Y) = 1 - \gamma$. We then define the data driven width of our confidence band:

$$\sigma_{n,\gamma} = \sigma_{n,\gamma}(Y) = \sup_{x\in[0,1]} \sum_{l=0}^{J_n} v_n \sqrt{\frac{\log n}{n}} \sum_{k=0}^{2^l-1} 1_{\{\tilde{f}_{lk}\neq 0\}} |\psi_{lk}(x)|, \tag{4.5}$$

where (v_n) is any (possibly data-driven) sequence such that $v_n \to \infty$. Under a local self-similarity type condition as in Kueh [28], one could possibly remove the supremum in (4.5) to obtain a spatially adaptive procedure. However, we restrict attention to more global self-similarity conditions here for simplicity. Since we consider wavelets satisfying (2.3), we have

$$\sigma_{n,\gamma} \le C(\psi) v_n \sqrt{\frac{\log n}{n}} \sum_{l=0}^{J_n} 2^{l/2} \le C' v_n \sqrt{\log n} < \infty \qquad \text{a.s.},$$

for all n and $\gamma \in (0,1)$. Let π_{med} denote the projection onto the nonzero coordinates of the posterior median and in a slight abuse of notation set

$$\pi_{\text{med}}(Y)(x) = \sum_{l=0}^{J_n} \sum_{k=0}^{2^l-1} Y_{lk} 1_{\{\tilde{f}_{lk} \neq 0\}} \psi_{lk}(x),$$

where we recall $Y_{lk} = \int_0^1 \psi_{lk}(t)\, dY(t)$. Consider the set

$$\overline{D}_n = \{ f : \|f - \mathbb{Y}\|_{\mathcal{M}(w)} \leq R_n/\sqrt{n}, \|f - \pi_{\text{med}}(Y)\|_\infty \leq \sigma_{n,\gamma}(Y) \}, \tag{4.6}$$

where R_n is as in (4.4). This involves a two-stage procedure: we firstly calculate the required $\mathcal{M}(w)$-radius R_n and then use the posterior median to select the coefficients deemed significant.

PROPOSITION 4.5. *Let $0 < \beta_1 \leq \beta_2 < \infty$, $R \geq 1$ and $\varepsilon > 0$. Consider the slab and spike prior defined above with threshold $j_0(n) \to \infty$ and let (w_l) be any admissible sequence that satisfies $w_{j_0(n)}/\sqrt{\log n} \nearrow \infty$. Then the confidence set $\overline{D}_n$ given in (4.6), using the choice (w_l) and $\sigma_{n,\gamma}(Y)$ defined in (4.5) for $v_n \to \infty$, satisfies*

$$\sup_{\substack{f_0 \in \mathcal{H}_{SS}(\beta,R,\varepsilon) \\ \beta \in [\beta_1,\beta_2]}} \left| \mathbb{P}_{f_0}(f_0 \in \overline{D}_n) - (1-\gamma) \right| \to 0$$

as $n \to \infty$. For every $\beta \in [\beta_1, \beta_2]$, uniformly over $f_0 \in \mathcal{H}_{SS}(\beta, R, \varepsilon)$,

$$\Pi(\overline{D}_n \mid Y) = 1 - \gamma + o_{\mathbb{P}_0}(1),$$

while the L^∞-diameter satisfies

$$|\overline{D}_n|_\infty = O_{\mathbb{P}_0}\big((n/\log n)^{-\beta/(2\beta+1)} v_n\big).$$

Under self-similarity, $\overline{D}_n$ has radius equal to the minimax rate in L^∞ up to some factor v_n that can be taken to diverge arbitrarily slowly, again mirroring a frequentist undersmoothing penalty. The choice of the posterior median is for simplicity and can be replaced by any other suitable thresholding procedure, for example directly using the posterior mixing probabilities between the atom at zero and the continuous density component.

One could also consider other alternatives to $\sigma_{n,\gamma}$ that simultaneously control the L^∞-norm of the credible set whilst also preserving coverage and credibility. A similar construction to the credible sets in Section 4.1 could also be pursued by intersecting D_n with a $B_{\infty 1}^{\hat{\beta}_n}$-ball, where $\hat{\beta}_n$ is a suitable estimate of the smoothness. Alternatively, in view of Remark 4.2, one can also show that

$$\Pi\left(f \in D_n : \|f - T_n\|_\infty \leq \left(\frac{w_{j_n(\beta)}}{\sqrt{j_n(\beta)}} n^{-\beta/(2\beta+1)} (\log n)^{(\beta+1)/(2\beta+1)} \right) \,\middle|\, Y \right)$$
$$= 1 - \gamma + o_{\mathbb{P}_0}(1),$$

where T_n is an efficient estimator of f_0 in $\mathcal{M}$ that is also rate-optimal in L^∞ (e.g., (8.5) or (8.6) in the Supplement [39]) and $2^{j_n} \sim (n \log n)^{1/(2\beta+1)}$. The factor $w_{j_n(\beta)}/\sqrt{j_n(\beta)}$ can be made to diverge arbitrarily slowly by the prior choice of $j_0(n)$.

5. Posterior independence of the credible sets. As shown above, the spaces $H(\delta) = H_2^{-1/2,\delta}$ and $\mathcal{M}(w)$ yield credible sets with good frequentist properties. However, given the different geometries proposed, it is of interest to compare them to more classical credible sets. Consider the ℓ_2-ball studied in [18, 43] (though without the blow-up factor of the latter)

$$C_n^{\ell_2} = \{f : \|f - \hat{f}_n\|_2 \le \tilde{Q}_n(\hat{\alpha}_n, \gamma)\}, \tag{5.1}$$

where $\tilde{Q}_n(\hat{\alpha}_n, \gamma)$ is selected such that $\Pi_{\hat{\alpha}_n}(C_n^{\ell_2} \mid Y) = 1 - \gamma$. Since the posterior variance of $\Pi_\alpha(\cdot \mid Y)$ is independent of the data, the radius $\tilde{Q}_n(\hat{\alpha}_n, \gamma)$ depends only on the data through $\hat{\alpha}_n$. By Theorem 1 of [18], we have $\tilde{Q}_n(\alpha, \gamma) = Q_n n^{-\alpha/(2\alpha+1)}$, where $Q_n \to Q > 0$.

Numerical examples of $\tilde{C}_n$ and $C_n^{\ell_2}$ are displayed in Section 6. Given the similarity of $\tilde{C}_n$ and $C_n^{\ell_2}$ in Figures 1 and 2, a natural question (voiced in [10, 34, 44]) is to what extent these sets actually differ, both in theory and practice. From a purely geometric point of view, these sets can be considered as infinite-dimensional ellipsoids with differing orientations. From a Bayesian perspective, an intriguing question is to what degree the decision rules on which these credible sets are based differ with respect to the posterior. For simplicity, we centre $\tilde{C}_n$ at the posterior mean $\hat{f}_n$, which we can do by Remark 4.3.

THEOREM 5.1. *Suppose a_n in (3.2) satisfies $a_n \le \log n/(6 \log n \log n)$. Then the $(1-\gamma)$-$H(\delta)$-credible ball $\tilde{C}_n$ defined in (4.2) and the $(1-\gamma)$-ℓ_2-credible ball $C_n^{\ell_2}$ defined in (5.1) are asymptotically independent under the empirical Bayes posterior, that is, as $n \to \infty$,*

$$\Pi_{\hat{\alpha}_n}(\tilde{C}_n \cap C_n^{\ell_2} \mid Y) = \Pi_{\hat{\alpha}_n}(\tilde{C}_n \mid Y)\Pi_{\hat{\alpha}_n}(C_n^{\ell_2} \mid Y) + o_{\mathbb{P}_0}(1) = (1-\gamma)^2 + o_{\mathbb{P}_0}(1)$$

uniformly over $f_0 \in \mathcal{Q}(\beta, R)$.

The first equality above also holds with $\tilde{C}_n$ replaced by C_n or $C_n^{\ell_2}$ replaced by the blown-up ℓ_2-credible ball studied in [43]. Moreover, the above statement also holds for the hierarchical Bayes posterior with $\tilde{C}_n$ replaced by the $(1-\gamma)$-$H(\delta)$-credible ball $\tilde{C}_n'$ given in (4.3) and $C_n^{\ell_2}$ replaced by the corresponding hierarchical Bayes ℓ_2-credible set.

Theorem 5.1 says that the Bayesian decision rules leading to the construction of $\tilde{C}_n$ and $C_n^{\ell_2}$ are fundamentally unrelated—one contains asymptotically no information about the other. Although we can conclude that $\tilde{C}_n$ and (blown-up) $C_n^{\ell_2}$ are

frequentist confidence sets with similar properties, they express completely different aspects of the posterior. Note that this is not simply an artefact of the prior choice since the equivalent prior credible sets are not independent under the prior despite its product structure. An alternative interpretation is to consider Bayesian tests based on the credible regions, which have optimal frequentist properties. In this context, the two tests screen different and unrelated features. While both of these approaches are valid, both for the frequentist and the Bayesian, Theorem 5.1 says that neither of these constructions can be reduced to the other.

The $H(\delta)$-credible sets are principally determined by the low frequencies ($k \leq k_n$, where $k_n \to \infty$ in the proof), whereas the ℓ_2-credible sets are driven by the high frequencies ($k > k_n$). The product structure of the posterior asymptotically decouples these two regimes yielding the independence statement. In particular, the $H(\delta)$-norm down-weights the higher order frequencies enough that one is dealing with a close to finite-dimensional model. Such a result is unlikely to hold for arbitrary priors, unless there is some degree of posterior independence between the frequency ranges driving the different credible sets (though less independence than a full product posterior is necessary). The numerical simulations in Table 1 corroborate Theorem 5.1 very closely, indicating that this result provides a good finite sample approximation to the posterior behaviour.

The posterior draws plotted in Section 6 are approximately drawn from the posterior distribution conditioned to the respective credible sets. Corollary 5.2 quantifies how close these draws are in terms of the total variation distance $\|\cdot\|_{\mathrm{TV}}$.

COROLLARY 5.2. *Let* $\Pi_{\hat{\alpha}_n}^{\tilde{C}_n}(\cdot \mid Y)$, $\Pi_{\hat{\alpha}_n}^{C_n^{\ell_2}}(\cdot \mid Y)$ *denote the posterior distribution conditioned to the sets* $\tilde{C}_n$, $C_n^{\ell_2}$, *respectively. Then as* $n \to \infty$,

$$\big\|\Pi_{\hat{\alpha}_n}^{\tilde{C}_n}(\cdot \mid Y) - \Pi_{\hat{\alpha}_n}^{C_n^{\ell_2}}(\cdot \mid Y)\big\|_{\mathrm{TV}} = \gamma + o_{\mathbb{P}_0}(1).$$

PROOF. Each conditional distribution consists of the posterior distribution restricted to the relevant credible set and normalized by the same factor $(1-\gamma)$. The two distributions are therefore identical on their intersection and so twice the total variation distance equals

$$\frac{\Pi_{\hat{\alpha}_n}(\tilde{C}_n \cap (C_n^{\ell_2})^c \mid Y)}{\Pi_{\hat{\alpha}_n}(\tilde{C}_n \mid Y)} + \frac{\Pi_{\hat{\alpha}_n}(\tilde{C}_n^c \cap C_n^{\ell_2} \mid Y)}{\Pi_{\hat{\alpha}_n}(C_n^{\ell_2} \mid Y)} = \frac{2\gamma(1-\gamma) + o_{\mathbb{P}_0}(1)}{1-\gamma}. \qquad \square$$

Turning to the L^∞-setting, for mathematical convenience let us consider the slightly stronger Besov norm $\|f\|_{B_{\infty 1}^0} = \sum_l 2^{l/2} \max_k |\langle f, \psi_{lk}\rangle|$ as in Hoffmann et al. [24]. This norm is closely related to the $\|\cdot\|_\infty$-norm via the Besov space embbedings $B_{\infty 1}^0 \subset L^\infty \subset B_{\infty\infty}^0$ (Chapter 4.3 of [22]). Define

$$D_n^{L^\infty} = \{f : \|f - T_n\|_{B_{\infty 1}^0} \leq \overline{Q}_n(\gamma)\}, \tag{5.2}$$

where $T_n = T_n^{(2)}$ is given by (8.6) in the Supplement [39] and is an efficient estimator of f_0 in $\mathcal{M}$ that is also rate-optimal in L^∞ and $\overline{Q}_n(\gamma)$ is selected such that $\Pi(D_n^{L^\infty} \mid Y) = 1 - \gamma$. The choice of T_n is not essential, but it is convenient to select an estimator that can simultaneously act as the centering for both D_n and $D_n^{L^\infty}$. We take the density g in Π to be Gaussian to simplify certain computations. Analogous results to those in $H(\delta)$ then hold.

THEOREM 5.3. *Consider the slab and spike prior* Π *with lower threshold* $j_0(n) \to \infty$ *and let* g *be the density of the Gaussian distribution* $N(0, \tau^2)$. *Let* (w_l) *be any admissible sequence satisfying* $w_{j_0(n)}/\sqrt{\log n} \nearrow \infty$ *as* $n \to \infty$. *Then the* $(1-\gamma)$*-*$\mathcal{M}(w)$*-credible ball* $\overline{D}_n$ *defined in* (4.6) *and the* $(1-\gamma)$*-*L^∞*-credible ball* $D_n^{L^\infty}$ *defined in* (5.2) *are asymptotically independent under the posterior, that is, as* $n \to \infty$,

$$\Pi\big(\overline{D}_n \cap D_n^{L^\infty} \mid Y\big) = \Pi(\overline{D}_n \mid Y)\Pi\big(D_n^{L^\infty} \mid Y\big) + o_{\mathbb{P}_0}(1) = (1-\gamma)^2 + o_{\mathbb{P}_0}(1)$$

uniformly over $f_0 \in \mathcal{H}(\beta, R)$.

In particular the choice of $j_0(n)$ in Corollary 3.6 satisfies the conditions of Theorem 5.3 since then $w_{j_0(n)} \simeq u_n\sqrt{\log n}$, where u_n can be made to diverge arbitrarily slowly.

COROLLARY 5.4. *Consider the same conditions as in Theorem* 5.3 *and let* $\Pi^{\overline{D}_n}(\cdot \mid Y)$, $\Pi^{D_n^{L^\infty}}(\cdot \mid Y)$ *denote the posterior distribution conditioned to the sets* $\overline{D}_n$, $D_n^{L^\infty}$ *respectively. Then as* $n \to \infty$,

$$\big\|\Pi^{\overline{D}_n}(\cdot \mid Y) - \Pi^{D_n^{L^\infty}}(\cdot \mid Y)\big\|_{\mathrm{TV}} = \gamma + o_{\mathbb{P}_0}(1).$$

Heuristics for an extension to density estimation. The proofs of Theorems 5.1 and 5.3 presented here rely on the independence of the coordinates in the Gaussian white noise model. This model can be viewed as an idealized version of other more concrete statistical models, being mathematically more tractable. In view of the extension of the nonparametric BvM to density estimation in [12], let us briefly discuss a heuristic of what we might expect in this setting.

Suppose we observe $Y_1, \dots, Y_n$ i.i.d. observations from an unknown density f_0 on $[0,1]$. Assume that f_0 is uniformly bounded away from 0 and that $f_0 \in C^\beta([0,1])$, where $1/2 < \beta \le 1$. Consider a simple histogram prior Π:

$$f = 2^L \sum_{k=0}^{2^L-1} h_k 1_{I_{Lk}}, \qquad I_{Lk} = \big(k2^{-L}, (k+1)2^{-L}\big], \qquad k \ge 0,$$

where the h_k are drawn from a $\mathcal{D}(1, \dots, 1)$-Dirichlet distribution on the unit simplex in $\mathbb{R}^{2^L}$. We ignore adaptation issues and select $L = L_n \to \infty$ based on the

smoothness of f_0. Such a prior has been shown to contract optimally in L^∞ by Castillo [9] and to satisfy a weak BvM in $\mathcal{M}_0$ by Castillo and Nickl [12].

Let (ψ_{lk}) denote the Haar wavelet basis on $[0,1]$ and $\mathcal{M}$ the related multiscale space for a suitable admissible sequence (w_l). Further let π_A denote the projection onto the elements of the Haar wavelet basis with resolution level contained in A. One can show that for suitable sequences $j_{n,1}, j_{n,2} \to \infty$ satisfying $2^{j_{n,1}} \ll 2^{j_{n,2}}$,

$$\begin{aligned}\Pi\big(f : \|f - T_n\|_{\mathcal{M}} \le R_n/\sqrt{n}, \|f - T_n\|_\infty \le \bar{Q}_n \mid Y\big) \\ = \Pi\big(f : \big\|\pi_{\le j_{n,1}}(f - T_n)\big\|_{\mathcal{M}} \le \big(R_n + o(1)\big)/\sqrt{n}, \\ \big\|\pi_{\ge j_{n,2}}(f - T_n)\big\|_\infty \le \bar{Q}_n + \delta_n | Y\big) + o_{\mathbb{P}_0}(1),\end{aligned}$$

where T_n is a suitable centering, $R_n/\sqrt{n}$ and $\bar{Q}_n$ are the $(1-\gamma)$-quantiles of the respective credible sets and δ_n is selected small enough to only change the credibility of the latter set by $o_{\mathbb{P}_0}(1)$.

Using the conjugacy of the Dirichlet distribution with multinomial sampling, the posterior distribution for (h_k) is $\mathcal{D}(N_1 + 1, \dots, N_L + 1)$, where $N_k = |\{Y_i : Y_i \in I_{Lk}\}|$. Observe that the law of $f_{lk} = \langle f, \psi_{lk}\rangle$ under the posterior depends principally on the observations falling within $\operatorname{supp}(\psi_{lk}) = [k2^{-l}, (k+1)2^{-l}]$. Unlike the Gaussian white noise model, there is dependence across the posterior wavelet coefficients due to the dependence within the Dirichlet distribution and the constraint that the number of observations sums to n.

The $\|\cdot\|_{\mathcal{M}}$-norm in the above display is the weighted maximum of $\{|f_{lk}| : l \le j_{n,1}\}$. If $j_{n,1}$ does not grow too fast, this consists of relatively few "large sample" averages. Heuristically, this term behaves like a central order statistic, being driven by the average sample behaviour. On the contrary, the $\|\cdot\|_\infty$-term is determined by the largest coefficients at each resolution level $l \ge j_{n,2}$. Since $2^{j_{n,1}} \ll 2^{j_{n,2}}$, these can be seen to behave more like extreme order statistics, being the maximum of many almost independent "small samples" (at least relative to the frequencies $l \le j_{n,1}$). Even though order statistics depend, by definition, on all observations, central and extreme order statistics asymptotically depend on the observations in orthogonal ways and become stochastically independent (cf. Chapter 21 of [45]). One might therefore hope that the two norms in the previous display are asymptotically independent in the sense of Theorems 5.1 and 5.3.

An alternative way to understand why the wavelet coefficients at a given resolution level may be considered "almost independent" under the posterior is via Poissonization. It is well known that density estimation is asymptotically equivalent to Poisson intensity estimation [32, 40], where one observes a Poisson process with intensity measure nf_0. Equivalently, the Poisson experiment corresponds to observing a Poisson random variable N with expectation n and then independently of N observing $Y_1, \dots, Y_N$ i.i.d. with density f_0. In this framework, the dependence induced by the number of observations summing to n is removed, meaning that the variables $(N_1, \dots, N_L)$ defined above are fully independent. The

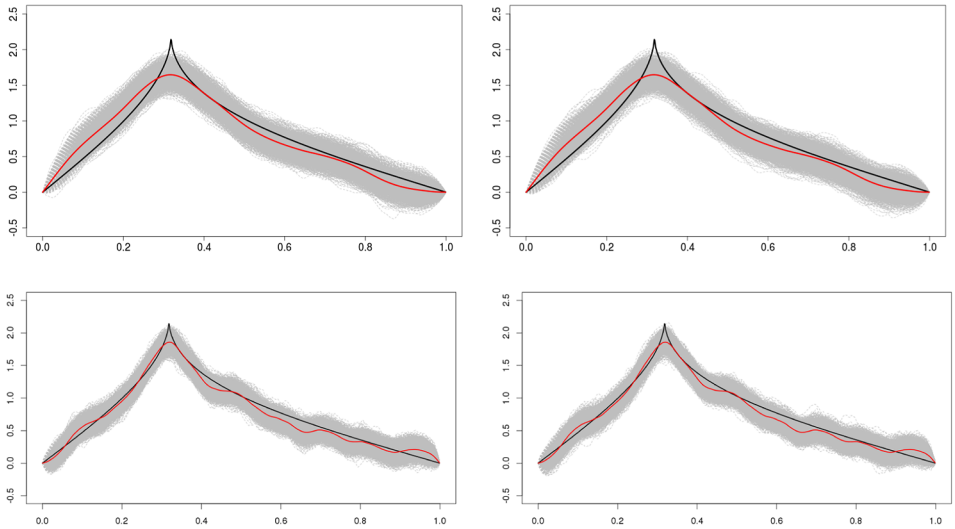

FIG. 1. *Empirical Bayes credible sets for the Fourier sine basis with the true curve (black) and the empirical Bayes posterior mean (red). The left panels contain the ℓ_2 credible ball $C_n^{\ell_2}$ given in (5.1) and the right panels contain the set $\tilde{C}_n$ given in (4.2). From top to bottom, $n = 500, 2000$ and $\hat{\alpha}_n = 1.29, 1.01$, with the right-hand side each having credibility 95%.*

remaining dependence is due to the Dirichlet distribution and becomes negligible as the number of bins $L_n \to \infty$. Since this equivalence is asymptotic in nature, one should expect such a heuristic to manifest itself also asymptotically.

6. Simulation example. We apply our approach in a numerical example, considering first the space $H_2^{-1/2,\delta}$. Consider the Fourier sine basis

$$e_k(x) = \sqrt{2}\sin(k\pi x), \qquad k = 1, 2, \ldots,$$

and define the true function $f_{0,k} = \langle f_0, e_k \rangle_2 = k^{-3/2}\sin(k)$ so that the true smoothness is $\beta = 1$. We consider realisations of the data (2.2) at levels $n = 500$ and 2000 and use the empirical Bayes posterior distribution. We plotted the true f_0 (black), the posterior mean (red) and an approximation to the credible sets (grey). To simulate the ℓ_2 credible balls $C_n^{\ell_2}$ given in (5.1), we sampled 2000 curves from the posterior distribution and kept the 95% closest in the ℓ_2 sense to the posterior mean and plotted them (grey). We performed the same approach to obtain the full $H(\delta)$-credible set C_n given in (4.1) and then plotted the full adaptive confidence set $\tilde{C}_n$ given in (4.2) with $C = 1$ and $\epsilon_n = 1/\log n$. We also present the approximate credibility of $\tilde{C}_n$ by considering the fraction of the simulated curves from the posterior that satisfy the extra constraint of $\tilde{C}_n$ that $\|f - \hat{f}_n\|_{H^{\hat{\alpha}_n - \epsilon_n}} \leq \sqrt{\log n}$. This is given in Figure 1.

While the true ℓ_2 and $H(\delta)$ credible balls are unbounded in L^∞, the posterior draws can be shown to be bounded in L^∞ explaining the boundedness of the plots.

TABLE 1
Table showing the average credibility of $\tilde{C}_n$, the average credibility of the posterior draws falling in both sets and the expected value of the latter (from Theorem 5.1)

		$n = 500$		
Chosen significance	0.95	0.90	0.85	0.80
Credibility of $\tilde{C}_n$	0.9500	0.8999	0.8499	0.8000
Credibility of $\tilde{C}_n \cap C_n^{\ell_2}$	0.9020	0.8102	0.7220	0.6406
Expected credibility of $\tilde{C}_n \cap C_n^{\ell_2}$	0.9025	0.8100	0.7225	0.6400
		$n = 2000$		
Chosen significance	0.95	0.90	0.85	0.80
Credibility of $\tilde{C}_n$	0.9500	0.9000	0.8500	0.8000
Credibility of $\tilde{C}_n \cap C_n^{\ell_2}$	0.9025	0.8095	0.7226	0.6409
Expected credibility of $\tilde{C}_n \cap C_n^{\ell_2}$	0.9025	0.8100	0.7225	0.6400

Sampling from the posterior (and thereby implicitly intersecting the sets $C_n^{\ell_2}$ and $\tilde{C}_n$ with the posterior support) seems the natural approach for the Bayesian. Indeed those elements that constitute the "roughest" or least regular elements of the credible sets are not seen by the posterior, that is, they have little or no posterior mass (see Lemma 8.3 in the Supplement [39]). The posterior contains significantly more information than merely the ℓ_2 or $H(\delta)$ norm of the parameter of interest, as can be seen by it assigning mass 1 to a strict subset of ℓ_2. For further discussion on plotting such credible sets, see [10, 31, 44].

For a given set of 2000 posterior draws, we also computed the credibility of $\tilde{C}_n$ at a chosen significance level and the credibility of the posterior draws falling in both $\tilde{C}_n$ and $C_n^{\ell_2}$. This latter quantity has value $(1-\gamma)^2 + o_{\mathbb{P}_0}(1)$ by Theorem 5.1. We repeated this 20 times and the average values are presented in Table 1.

The posterior distribution appears to have some difficulty visually capturing the resulting function at its peak. In fact, the credible sets do "cover the true function", but do so in an ℓ_2 rather than an L^∞-sense. Indeed, any ℓ_2-type confidence ball will be unresponsive to highly localized pointwise features since they occur on a set of small Lebesgue measure (as in this case). Similar reasoning also explains the performance of the posterior mean at this point. The posterior mean estimates the Fourier coefficients of f_0, and hence estimates the true function in an ℓ_2-sense via its Fourier series.

In Section 5, it was shown that the two approaches behave very differently theoretically, and the numerical results in Table 1 match this theory very closely. It appears that the two methods do indeed use different rejection criteria in practice resulting in different selection outcomes. The visual similarity between the ℓ_2 and $H(\delta)$-credible balls in Figure 1 is therefore a result of the posterior draws themselves looking similar, rather than the methods performing identically.

We note that already by $n = 500$, $\tilde{C}_n$ has the correct credibility so that the high frequency smoothness constraint is satisfied with posterior probability virtually

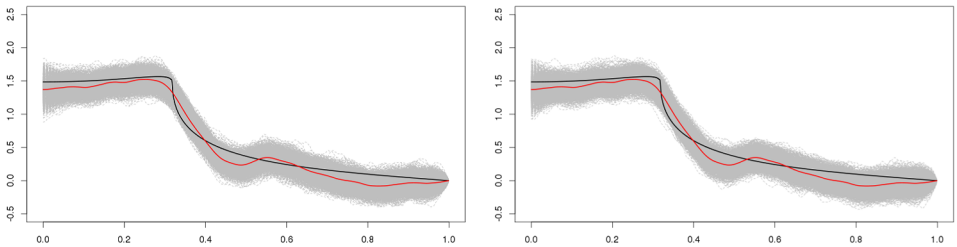

FIG. 2. *Empirical Bayes credible sets for the Volterra SVD basis with the true curve (black) and the empirical Bayes posterior mean (red) for* $n = 1000$ *and* $\hat{\alpha}_n = 1.07$. *The left and right panels contain the* ℓ_2 *credible ball* $C_n^{\ell_2}$ *given in* (5.1) *and* $\tilde{C}_n$ (*credibility* 95%) *given in* (4.2), *respectively.*

equal to one (cf. Proposition 4.1). $\tilde{C}_n$ is therefore an actual credible set for reasonable (finite) sample sizes rather than a purely asymptotic credible set. The posterior distribution already strongly regularizes the high frequencies so that the posterior draws are very regular with high probability. This can be quantitatively seen by the rapidly decaying variance term of the posterior distribution (3.1). This is indeed the case in the simulation, where the credibility gap is negligible, thereby demonstrating that most of the posterior draws already satisfy the smoothness constraint in $\tilde{C}_n$.

We repeat the same simulation using the same true function $f_{0,k} = k^{-3/2}\sin(k)$, but with basis equal to the singular value decomposition (SVD) of the Volterra operator (cf. [27]):

$$e_k(x) = \sqrt{2}\cos\bigl((k - 1/2)\pi x\bigr), \qquad k = 1, 2, \ldots$$

and plot this in Figure 2 for $n = 1000$. Unlike Figure 1, the resulting function has no "spike" and so both credible sets have no trouble visually capturing the true function (though one should remember that these are ℓ_2 rather than L^∞ type credible sets).

We now illustrate the multiscale approach using the slab and spike prior with lower threshold $j_0(n) = \sqrt{\log n}$, plotting the true function (solid black) and posterior mean (red) at levels $n = 200, 500$. We have used Haar wavelets, set g to be $N(0, 1/2)$ and have taken prior weights $w_{j,n} = \min(n^{-1}, 2^{-5.5j})$, corresponding to $K = 1$ and $\theta = 4.5$. For $n = 200$ and 500, we have fitted one scaling function plus $2^8 - 1 = 255$ and $2^9 - 1 = 511$ wavelet coefficients respectively (i.e., $2^{J_n+1} - 1$). We again sampled 2000 curves from the posterior distribution and plotted the 95% closest to the posterior mean in the $\mathcal{M}(w)$ sense (grey) to simulate D_n in (4.4). We also used the posterior draws to generate a 95% credible band in L^∞ by estimating $\bar{Q}_n(0.05)$ and then plotting $D_n^{L^\infty}$ in (5.2) (dashed black). Finally, we computed local 95% credible intervals at every point $x \in [0, 1]$ and joined these to form a credible band (dashed blue). This is given in Figure 3.

We see from Figure 3 that each posterior draw consists of a rough approximation of the signal via frequencies $j \le j_0(n)$ with a few "spikes" from the high

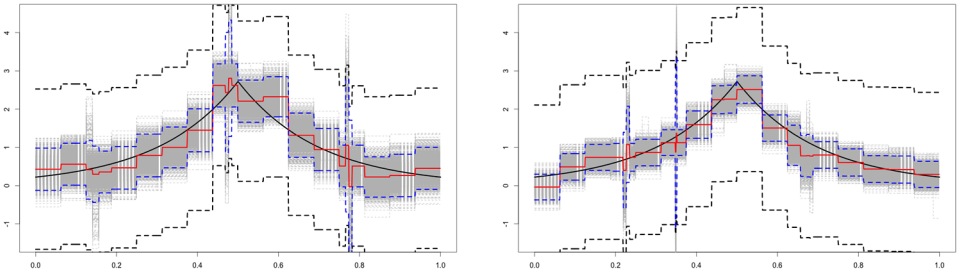

FIG. 3. *Slab and spike credible sets with the true curve* (*black*), *posterior mean* (*red*), *a* 95% *credible band in* L^∞ (*dashed black*), *pointwise* 95% *credible intervals* (*dashed blue*) *and the set* D_n *given in* (4.4) (*grey*). *We have* $n = 200, 500$, *respectively.*

frequencies; the rather unusual shape is a reflection of the prior choice. It is worth noting that the posterior draws are bounded in L^∞ since the posterior contracts rate optimally to the truth in L^∞ [24]. We see that the L^∞ diameter of D_n is strictly greater than that of the L^∞-credible bands, though this only manifests itself in a few places. The size of the L^∞-bands is driven by the size of the spikes, which are few in a number but occur in every posterior draw, resulting in seemingly very wide credible bands.

On the contrary, the local credible intervals ignore the spikes since less than 5% of the draws have a spike at any given point, resulting in much tighter bands. The dashed blue lines in effect correspond to the 95% L^∞-band from a prior fitting exclusively the low frequencies $j \leq j_0(n)$, which is a nonadaptive prior modelling analytic smoothness. This dramatically oversmoothes the truth resulting in far too narrow credible bands and is highly dangerous since it is known that oversmoothing the truth can yield zero coverage [27, 29].

Acknowledgements. The author would like to thank Richard Nickl, Aad van der Vaart, Johannes Schmidt-Hieber, the Associate Editor and two referees for their valuable comments. The author would like to express particular thanks to one referee for a very detailed report, including suggesting a simplified argument for Theorems 7.2 and 7.4 in the Supplement. Most of this work was completed during the author's PhD at the University of Cambridge.

SUPPLEMENTARY MATERIAL

Supplement to "Adaptive Bernstein–von Mises theorems in Gaussian white noise" (DOI: 10.1214/16-AOS1533SUPP; .pdf). All proofs together with additional results are given in the Supplement [39].

REFERENCES

[1] ABRAMOVICH, F., SAPATINAS, T. and SILVERMAN, B. W. (1998). Wavelet thresholding via a Bayesian approach. *J. R. Stat. Soc. Ser. B. Stat. Methodol.* **60** 725–749. MR1649547

[2] BELITSER, E. and GHOSAL, S. (2003). Adaptive Bayesian inference on the mean of an infinite-dimensional normal distribution. *Ann. Statist.* **31** 536–559. MR1983541
[3] BICKEL, P. J. and KLEIJN, B. J. K. (2012). The semiparametric Bernstein–von Mises theorem. *Ann. Statist.* **40** 206–237. MR3013185
[4] BONTEMPS, D. (2011). Bernstein–von Mises theorems for Gaussian regression with increasing number of regressors. *Ann. Statist.* **39** 2557–2584. MR2906878
[5] BOUCHERON, S. and GASSIAT, E. (2009). A Bernstein–von Mises theorem for discrete probability distributions. *Electron. J. Stat.* **3** 114–148. MR2471588
[6] BULL, A. D. (2012). Honest adaptive confidence bands and self-similar functions. *Electron. J. Stat.* **6** 1490–1516. MR2988456
[7] CASTILLO, I. (2008). Lower bounds for posterior rates with Gaussian process priors. *Electron. J. Stat.* **2** 1281–1299. MR2471287
[8] CASTILLO, I. (2012). A semiparametric Bernstein–von Mises theorem for Gaussian process priors. *Probab. Theory Related Fields* **152** 53–99.
[9] CASTILLO, I. (2014). On Bayesian supremum norm contraction rates. *Ann. Statist.* **42** 2058–2091. MR3262477
[10] CASTILLO, I. (2015). Discussion of "Frequentist coverage of adaptive nonparametric Bayesian credible sets". *Ann. Statist.* **43** 1437–1443. MR3357863
[11] CASTILLO, I. and NICKL, R. (2013). Nonparametric Bernstein–von Mises theorems in Gaussian white noise. *Ann. Statist.* **41** 1999–2028.
[12] CASTILLO, I. and NICKL, R. (2014). On the Bernstein–von Mises phenomenon for nonparametric Bayes procedures. *Ann. Statist.* **42** 1941–1969. MR3262473
[13] CASTILLO, I. and ROUSSEAU, J. (2015). A Bernstein–von Mises theorem for smooth functionals in semiparametric models. *Ann. Statist.* **43** 2353–2383.
[14] CASTILLO, I., SCHMIDT-HIEBER, J. and VAN DER VAART, A. (2015). Bayesian linear regression with sparse priors. *Ann. Statist.* **43** 1986–2018. MR3375874
[15] CASTILLO, I. and VAN DER VAART, A. (2012). Needles and straw in a haystack: Posterior concentration for possibly sparse sequences. *Ann. Statist.* **40** 2069–2101. MR3059077
[16] COX, D. D. (1993). An analysis of Bayesian inference for nonparametric regression. *Ann. Statist.* **21** 903–923. MR1232525
[17] DIACONIS, P. and FREEDMAN, D. (1986). On the consistency of Bayes estimates. *Ann. Statist.* **14** 1–67. MR0829555
[18] FREEDMAN, D. (1999). On the Bernstein–von Mises theorem with infinite-dimensional parameters. *Ann. Statist.* **27** 1119–1140. MR1740119
[19] GHOSAL, S. (1999). Asymptotic normality of posterior distributions in high-dimensional linear models. *Bernoulli* **5** 315–331.
[20] GINÉ, E. and NICKL, R. (2009). Uniform limit theorems for wavelet density estimators. *Ann. Probab.* **37** 1605–1646.
[21] GINÉ, E. and NICKL, R. (2010). Confidence bands in density estimation. *Ann. Statist.* **38** 1122–1170.
[22] GINÉ, E. and NICKL, R. (2015). *Mathematical Foundations of Infinite-Dimensional Statistical Models*. *Cambridge Series in Statistical and Probabilistic Mathematics* **40**. Cambridge Univ. Press, Cambridge.
[23] HOFFMANN, M. and NICKL, R. (2011). On adaptive inference and confidence bands. *Ann. Statist.* **39** 2383–2409.
[24] HOFFMANN, M., ROUSSEAU, J. and SCHMIDT-HIEBER, J. (2015). On adaptive posterior concentration rates. *Ann. Statist.* **43** 2259–2295.
[25] JOHNSTONE, I. M. (2010). High dimensional Bernstein–von Mises: Simple examples. In *Borrowing Strength*: *Theory Powering Applications—a Festschrift for Lawrence D. Brown*. *Inst. Math. Stat. Collect.* **6** 87–98. IMS, Beachwood, OH.

[26] KNAPIK, B. T., SZABÓ, B. T., VAN DER VAART, A. W. and VAN ZANTEN, J. H. (2016). Bayes procedures for adaptive inference in inverse problems for the white noise model. *Probab. Theory Related Fields* **164** 771–813. MR3477780

[27] KNAPIK, B. T., VAN DER VAART, A. W. and VAN ZANTEN, J. H. (2011). Bayesian inverse problems with Gaussian priors. *Ann. Statist.* **39** 2626–2657. MR2906881

[28] KUEH, A. (2012). Locally adaptive density estimation on the unit sphere using needlets. *Constr. Approx.* **36** 433–458.

[29] LEAHU, H. (2011). On the Bernstein–von Mises phenomenon in the Gaussian white noise model. *Electron. J. Stat.* **5** 373–404. MR2802048

[30] LE CAM, L. (1986). *Asymptotic Methods in Statistical Decision Theory*. *Springer Series in Statistics*. Springer, New York. MR0856411

[31] LOW, M. G. and MA, Z. (2015). Discussion of "Frequentist coverage of adaptive nonparametric Bayesian credible sets". *Ann. Statist.* **43** 1448–1454. MR3357865

[32] LOW, M. G. and ZHOU, H. H. (2007). A complement to Le Cam's theorem. *Ann. Statist.* **35** 1146–1165. MR2341701

[33] MEYER, Y. (1992). *Wavelets and Operators*. *Cambridge Studies in Advanced Mathematics* **37**. Cambridge Univ. Press, Cambridge.

[34] NICKL, R. (2015). Discussion of "Frequentist coverage of adaptive nonparametric Bayesian credible sets." *Ann. Statist.* **43** 1429–1436.

[35] NICKL, R. and SZABÓ, B. (2016). A sharp adaptive confidence ball for self-similar functions. *Stochastic Process. Appl.* **126** 3913–3934.

[36] PETRONE, S., ROUSSEAU, J. and SCRICCIOLO, C. (2014). Bayes and empirical Bayes: Do they merge? *Biometrika* **101** 285–302.

[37] RAY, K. (2013). Bayesian inverse problems with non-conjugate priors. *Electron. J. Stat.* **7** 2516–2549. MR3117105

[38] RAY, K. (2014). Asymptotic theory for Bayesian nonparametric procedures in inverse problems Ph.D. thesis, University of Cambridge.

[39] RAY, K. (2016). Supplement to "Adaptive Bernstein–von Mises theorems in Gaussian white noise". DOI:10.1214/16-AOS1533SUPP.

[40] RAY, K. and SCHMIDT-HIEBER, J. (2016). The Le Cam distance between density estimation, Poisson processes and Gaussian white noise. ArXiv E-prints.

[41] RIVOIRARD, V. and ROUSSEAU, J. (2012). Bernstein–von Mises theorem for linear functionals of the density. *Ann. Statist.* **40** 1489–1523.

[42] SZABÓ, B., VAN DER VAART, A. and VAN ZANTEN, H. (2015). Honest Bayesian confidence sets for the L^2-norm. *J. Statist. Plann. Inference* **166** 36–51.

[43] SZABÓ, B., VAN DER VAART, A. W. and VAN ZANTEN, J. H. (2015). Frequentist coverage of adaptive nonparametric Bayesian credible sets. *Ann. Statist.* **43** 1391–1428.

[44] SZABÓ, B., VAN DER VAART, A. W. and VAN ZANTEN, J. H. (2015). Rejoinder to discussions of "Frequentist coverage of adaptive nonparametric Bayesian credible sets". *Ann. Statist.* **43** 1463–1470. MR3357867

[45] VAN DER VAART, A. W. (1998). *Asymptotic Statistics*. *Cambridge Series in Statistical and Probabilistic Mathematics* **3**. Cambridge Univ. Press, Cambridge. MR1652247

MATHEMATICAL INSTITUTE
LEIDEN UNIVERSITY
NIELS BOHRWEG 1
2333 CA LEIDEN
THE NETHERLANDS
E-MAIL: k.m.ray@math.leidenuniv.nl